

**Project Title**

**Management of the Generation CP Central Registry**

**Targeted Subprogram**

Subprogram 4 (Bioinformatics)

**Principal Investigator and Lead Institution**

Tom Hazekamp (IPGRI)

**Collaborating Scientists and Institutions**

Marco Bink (WUR)

Subhash Chandra (ICRISAT)

Guy Davenport (CIMMYT)

Samy Gaiji (IPGRI)

Reinhard Simon (CIP)

Milko Skofic (IPGRI)

Rajesh Sood (IPGRI)

Dag Terje Endresen (IPGRI/NGB)

**Executive Summary**

A large amount of data is being generated within the Generation Challenge Program. These data are stored and maintained at different locations, using different formats and standards. Organizing and publishing information on the Web through a Central Registry provides an overview of available data resources ('yellow pages' directory) from a single point. This is critical for the successful completion of the tasks that require data from various sources. The GCP Central Registry was established in 2005. In 2006 the aim of this project is to increase the depth and range of the resources it manages. The approach is to strengthen and further develop components of the Central Registry and to actively approach and assist GCP Partners to register new resources. The project focuses on components such as:

- The technical maintenance and management of the Central Registry
- To build up the Central Registry's resource collection through a pro-active approach of potential providers
- The further development of the Central Registry with indexing systems, data visualization tools and links to data analysis tools
- Content management resulting in more extensive controlled vocabularies and enhanced data validation rules
- Help desk to support providers and users

To obtain an insight in how users perceive the Central Registry, a selected group of users will be asked to provide feedback on the Central Registry's impact and the measure in which it meets their needs.

**Project Description**

**Rationale and Objectives**

The partners in the Generation Challenge Program are generating a large amount of data. These data are stored and maintained at different locations, using different methods and standards. Most importantly, these data are not available for all partners to use. Organizing and publishing information on the Web through a GCP Central Registry is a means to provide an overview of available data resources ('yellow pages' directory) from a single point. This is critical for the successful completion of the tasks that require

data from various sources. The GCP Central Registry was established in 2005. In 2006 the aim of this project is to increase the depth and range of the resources it manages.

## **Outputs/products**

### **Technical maintenance and management**

- Documented management procedures for the Central Registry
- Day-to-day management of Central Registry

### **Building up the Central Registry's resource collection**

- Central Registry providing a comprehensive overview of GCP produced resources

### **Development of the Central Registry**

- Indexing system for selected unit level data maintained by Central Registry
- Implementation of visualization tools to assist in the discovery of indexed unit-level data

### **Content management**

- Controlled vocabularies for types of meta, registration and unit level data
- Enhanced validation rules for registration data and unit level data

### **Help desk**

- Documented guidelines for data submission and use of the Central Registry
- Helpdesk established

### **Impact and User satisfaction**

- Report on Central Registry impact and user satisfaction assessment

## **Approach and methodology**

In 2006 the aim of this project is to increase the depth and width of the overview that the Central Registry provides of GCP resources. The approach is to strengthen and further develop components of the Central Registry and to actively approach and assist GCP Partners to register new resources. The approach to achieve this will have the following components:

### ***1) Technical maintenance and management of the Central Registry***

Maintenance and management regimes for the Central Registry will be operated to ensure proper day-to-day operation. This requires the implementation of proper procedures for backup, virus protection etc. and a standing capacity to troubleshoot acute problems. Maintenance procedures include a strategy and implementation for the timely upgrade of hard- and software and Internet connectivity.

The server on which the Central Registry is developed and run is located at IPGRI Headquarters Rome. The technical maintenance and management of the Central Registry will be done by IPGRI Rome.

### ***2) Building up the Central Registry's resource collection***

#### ***Pro-active approach to stimulate submission of new resources.***

It is essential to increase the collection of resources held by the Central Registry for it to provide a comprehensive overview of resources produced by the GCP. Pro-active efforts to approach GCP partners to register resources were initiated during the last quarter of 2005 when the Central Registry was established. These efforts will be continued in 2006. It will entail communication with GCP partners holding resources, to assist them getting the resources committed to the Central Registry. This will include providing assistance with the conversion of data and the uploading of resources. These efforts will continue during the year, but the main thrust of efforts will be scheduled twice a year (Q1 and Q3) to establish a regular submission pattern. The registration of resources will generate metadata on past and currently produced resources by GCP projects. Jointly with the Data Template task, metadata will also be compiled on the types of data that are expected to be generated in the near-future by GCP projects. This data will on the one hand guide priority setting for data template development while on the other hand it provides the central registry with an insight in where new datasets will mature in the near future.

IPGRI will systematically approach of GCP partners to submit new resources to the Central Registry. The collection of metadata on future datasets will be jointly coordinated with the Data Template task led by CIMMYT.

### ***Downstream new standards developed within GCP***

To increase the range of resources managed the standards developed by other GCP projects need to be down streamed to providers as soon as they become available so that more data types can be submitted to the Central Registry in a standardized manner. This means that e.g. new data templates/domain models will be advertised and made available. Users will be assisted in the use of these new standards.

CIMMYT is leading the data template development. IRRI is leading the Domain modelling task in SP4 which will become more and more the focus for the development of GCP-wide standards. Coordination with both tasks is important to phase new standards in at the appropriate time and form.

### ***3) Development of the Central Registry***

New functionality will be developed ensure that the Central Registry continues to meet evolving user needs. The 'yellow page' directory services of the Central Registry providing an overview of information resources at the meta data level (i.e. describing the resources) is one of the core functions of the Central Registry. In addition services will be developed to provide a more detailed insight into resources registered. From datasets that are made available in XML format using GCP data templates or BioCASE webservices, core data items (e.g. taxonomic, location etc data) could be indexed by the Central Registry to provide a more detailed insight in the datasets. This requires the development of appropriate data harvesting and index systems. Which elements should be included as indexed data in the core set will need to be discussed with user and provider groups. The unit level data that will be indexed will require the development of new visualization tools to facilitate the discovery of this type of data.

The Central Registry will also provide linkages to and assessments of analysis tools that can assist further in the use of datasets. This will entail linkage and obtaining inputs from other SP4 projects such as Data analysis support for existing projects in SP1 with emphasis on sampling germplasm (WUR), Data analysis support for existing projects in SP2 with emphasis on integrating results from microarray and mapping experiments (CIP) and Development of decision support tools for MAS and MAB (ICRISAT).

The indexing system will be implemented by IPGRI as they are involved in the further technical maintenance of the system. To determine the core data types to build appropriate indexes, inputs will be sought from the lead institutes that develop templates for the various data types within the data template (led by CIMMYT). For the visualization tools CIP and IPGRI will work closely together. CIP has demonstrated experience with this type of tools (Mondrian, Datamart).

For inputs on analysis tools for SP1, SP2 and SP3, the institutes that lead tasks related to this area WUR, CIP and ICRISAT will be asked to provide assessments of suitable tools. In the case of ICRISAT some of these activities were already started in 2005.

CIMMYT will be tasked with the further development of XML parsers for conversion of XML formats to other output formats requested by users.

### ***4) Content management***

The data (meta data and other registration data) submitted to the Central Registry requires curation to ensure their efficient use. Quality standards and data validation rules will need to be further developed and applied (e.g. controlled vocabularies are necessary to facilitate the use of the Central Registry). When unit level data are to be included in index systems developed by the Central Registry, additional content management tasks will be added to ensure that unit level data can be efficiently accessed and used.

It is expected that the Central Registry could play an important role in the GCP Quality Assurance Strategy (to be developed as part of Improvement of Quality of Existing Databases task led by IRRI). The Central Registry would provide a good access point to GCP-wide datasets and could validate these resources against (to be) established quality assurance regulations.

IPGRI will lead the content management. It hosts the Central Registry databases and has experience in developing e.g. controlled vocabularies through its descriptor development activities over the years. In addition inputs will be sought from the lead institutes that develop data standards for the various data types to develop and or implement the required controlled vocabularies and validation rules.

**5) Helpdesk, manuals and guidelines**

The procedures for submission of resources will be documented to assist providers and users. Providers and users will require support at different levels. In addition to manuals and guidelines, a responsive helpdesk function needs to be put into place to provide real-time assistance.

IPGRI's role will be to develop the appropriate documentation and provide helpdesk support.

**6) Impact and user satisfaction**

In addition to the feedback received through the help desk functions, active feedback will be sought from user groups to assess the type of impact the Central Registry has and to ensure that there exists a clear and up-to-date picture of what the various user groups within the GCP expect from the Central Registry and that those expectations are adequately met.

Representatives from the various user groups will be asked to actively comment how the Central Registry impacts on their activities and how it responds to their needs. The findings will be reported to GCP SP4 Leadership.

**Partners (and their role in the project)**

Institute	Collaborator	Role
CIMMYT	Guy Davenport	- Inputs on further development of Central Registry vis-à-vis standards coming out of Data template task - Development of XML converters to generate specific outputs format from XML formatted files to other formats
CIP	Reinhard Simon	- Assessment and implementation of visualization tools for Central Registry - Assessments of data analysis tools for SP2
ICRISAT	Subhash Chandra	- Assessments of data analysis tools for SP3
WUR	Marco Bink	- Assessments of data analysis tools for SP1

## Timeline and Milestones

Components	Q1	Q2	Q3	Q4
<b>1) Technical maintenance and management</b>				
Development of management documentation				
Technical maintenance and management				
<b>2) Building up the Central Registry's resource collection</b>				
Pro-active approach to providers				
Downstream standards (as they become avail.)				
<b>3) Development of Central Registry</b>				
Data indexing (as templates are developed)				
Visualization tools				
Assessments analysis tools				
XML converters				
<b>4) Content management</b>				
Developing CV's, update validation rules				
<b>5) Helpdesk manuals and guidelines</b>				
Development manuals and guidelines				
Helpdesk				
<b>6) Impact and User satisfaction</b>				
User group consultation and report				

## Linkages with Other Projects (within GCP and outside)

- Development of Generation CP domain models (IRRI)
- Implementation of web services technology in Generation CP Consortium (IPGRI)
- Creation and maintenance of templates for the GCP data storage in repositories (CIMMYT)
- Improvement of quality of existing databases (IRRI)
- Data analysis support for existing projects in SP1 with emphasis on sampling germplasm (WUR)
- Data analysis support for existing projects in SP2 with emphasis on integrating results from microarray and mapping experiments (CIP)
- Development of decision support tools for MAS and MAB (ICRISAT)

## Critical Assumptions and Contingency Plans

The provision of new standards for additional data types from the Data Template or Domain Modelling Tasks is critical for the Central Registry to extend its scope of resources managed. Without additional data standards, it will still be possible for the Central Registry to record information at the metadata level on these datasets, but it will not be possible to provide a unit-level insight in these datasets. Also without common standards in place and applied users will experience additional difficulties to analyse data that span across datasets.

## Budget

### Budget by Partner by Year (in US\$)

LEAD INSTITUTION (IPGRI)	Year 1 & Total
Personnel costs	61.500
Supplies and services	
Field	
Lab	7.500
Travel	7.500
Training, meeting, and workshop	
<b>Subtotal</b>	<b>76.500</b>
Indirect costs (18%)	13.770

<b>Lead Institution Total</b>	<b>90.270</b>
<b>PARTNER 2 (CIMMYT)</b>	<b>Year 1 &amp; Total</b>
Personnel costs	9.000
Supplies and services	
Field	
Lab	
Travel	
Training, meeting, and workshop	
<b>Subtotal</b>	<b>9.000</b>
Indirect costs (18%)	1.620
<b>Partner 2 Total</b>	<b>10.620</b>
<b>PARTNER 3 (CIP)</b>	<b>Year 1 &amp; Total</b>
Personnel costs	9.000
Supplies and services	
Field	
Lab	
Travel	
Training, meeting, and workshop	
<b>Subtotal</b>	<b>9.000</b>
Indirect costs (18%)	1.620
<b>Partner 3 Total</b>	<b>10.620</b>
<b>PARTNER 4 (ICRISAT)</b>	<b>Year 1 &amp; Total</b>
Personnel costs	3.000
Supplies and services	
Field	
Lab	
Travel	
Training, meeting, and workshop	
<b>Subtotal</b>	<b>3.000</b>
Indirect costs (18%)	540
<b>Partner 4 Total</b>	<b>3.540</b>
<b>PARTNER 5 (WUR)</b>	<b>Year 1 &amp; Total</b>
Personnel costs	5.000
Supplies and services	
Field	
Lab	
Travel	
Training, meeting, and workshop	
<b>Subtotal</b>	<b>5.000</b>
Indirect costs (18%)	900
<b>Partner 5 Total</b>	<b>5.900</b>
<b>GRAND TOTAL</b>	<b>120.950</b>

## Budgets per activity

### Coordination

	IPGRI	CIMMYT	CIP	ICRISAT	WUR	Total
Personnel costs	15.000					15.000
Supplies and services						
Field						
Lab						
Travel	7.500					7.500
Training, meeting, and workshop						
<b>Subtotal</b>	<b>22.500</b>					<b>22.500</b>
Indirect costs (18%)	4.050					4.050
<b>Total</b>	<b>26.550</b>					<b>26.550</b>

### Technical maintenance and management

	IPGRI	CIMMYT	CIP	ICRISAT	WUR	Total
Personnel costs	7.000					7.000
Supplies and services						
Field						
Lab	7.500					7.500
Travel						
Training, meeting, and workshop						
<b>Subtotal</b>	<b>14.500</b>					<b>14.500</b>
Indirect costs (18%)	2.610					2.610
<b>Total</b>	<b>17.110</b>					<b>17.110</b>

### Building up the Central Registry's resource collection

	IPGRI	CIMMYT	CIP	ICRISAT	WUR	Total
Personnel costs	18.275					18.275
Supplies and services						
Field						
Lab						
Travel						
Training, meeting, and workshop						
<b>Subtotal</b>	<b>18.275</b>					<b>18.275</b>
Indirect costs (18%)	3.290					3.290
<b>Total</b>	<b>21.565</b>					<b>21.565</b>

### Development of the Central Registry

	IPGRI	CIMMYT	CIP	ICRISAT	WUR	Total
Personnel costs	11.650	9.000	9.000	3.000	5.000	37.650
Supplies and services						
Field						
Lab						
Travel						
Training, meeting, and workshop						
<b>Subtotal</b>	<b>11.650</b>	<b>9.000</b>	<b>9.000</b>	<b>3.000</b>	<b>5.000</b>	<b>37.650</b>

Indirect costs (18%)	2.100	1.620	1.620	540	900	<b>6.780</b>
<b>Total</b>	<b>13.750</b>	<b>10.620</b>	<b>10.620</b>	<b>3.540</b>	<b>5.900</b>	<b>44.430</b>

### Content management

	IPGRI	CIMMYT	CIP	ICRISAT	WUR	Total
Personnel costs	2.650					<b>2.650</b>
Supplies and services						
Field						
Lab						
Travel						
Training, meeting, and workshop						
<b>Subtotal</b>	<b>2.650</b>					<b>2.650</b>
Indirect costs (18%)	419					<b>419</b>
<b>Total</b>	<b>3.069</b>					<b>3.069</b>

### Help desk

	IPGRI	CIMMYT	CIP	ICRISAT	WUR	Total
Personnel costs	4.650					<b>4.650</b>
Supplies and services						
Field						
Lab						
Travel						
Training, meeting, and workshop						
<b>Subtotal</b>	<b>4.650</b>					<b>4.650</b>
Indirect costs (18%)	837					<b>837</b>
<b>Total</b>	<b>5.487</b>					<b>5.487</b>

### Impact and User satisfaction

	IPGRI	CIMMYT	CIP	ICRISAT	WUR	Total
Personnel costs	2.325					<b>2.325</b>
Supplies and services						
Field						
Lab						
Travel						
Training, meeting, and workshop						
<b>Subtotal</b>	<b>2.325</b>					<b>2.325</b>
Indirect costs (18%)	418					<b>418</b>
<b>Total</b>	<b>2.743</b>					<b>2.743</b>

### Budget Notes and Justification

#### IPGRI

The personnel cost for IPGRI (\$61.500) cover project coordination (10% IRS and 10% programme support for \$15.000), technical support staff (\$7.000) and staff/consultancy fees (\$39.500) to execute the various project activities.

The travel budget (\$7.500) will be used to bring in or visit collaborators on technical aspects of the Central Registry such as CIMMYT or CIP for face-to-face discussions on implementation aspects. It can also be used to strengthen linkages with other projects by attending one of their meetings.

The budget for supplies (\$7.500) will be used to upgrade Central Registry hard-/software and upgrade internet connectivity as necessary.

**CIMMYT**

The personnel cost for CIMMYT (\$9.000) is to cover development of converters that allow XML formatted data to be transformed in other formats required by users. These could include e.g. transformation of XML to MS Excel or Access formats.

Furthermore inputs are expected into the development of the Central Registry in particular to make sure that the GCP Data Templates and the Central Registry are well aligned.

**CIP**

The personnel cost for CIP (\$9.000) is to cover assistance with the implementation of visualization tools for the Central Registry.

Furthermore written assessments are expected on data analysis tools for SP2. The write-up for particular analysis tools are expected to cover aspects such as the purpose for which the tool can be used, how it can be obtained, its performance, strong-weak points etc.

**WUR**

The personnel cost for WUR (\$5.000) is to cover written assessments on data analysis tools for SP1. The write-up for particular analysis tools are expected to cover aspects such as the purpose for which the tool can be used, how it can be obtained, its performance, strong-weak points etc.

**ICRISAT**

The personnel cost for ICRISAT (\$3.000) is to cover written assessments on data analysis tools for SP3. The write-up for particular analysis tools are expected to cover aspects such as the purpose for which the tool can be used, how it can be obtained, its performance, strong-weak points etc. This is a continuation of the work started in 2005.